

The Planets Testbed: A collaborative environment for experimentation in digital preservation

Sven Schlarb
Austrian National Library
Josefsplatz 1, A-1015 Vienna, Austria
Tel.: +43-1-53410-491, Email: sven.schlarb@onb.ac.at

Andrew N. Jackson
The British Library
Boston Spa, West Yorkshire, LS23 7BQ, UK
Tel.: +44-1937-546602, Email: andrew.jackson@bl.uk

Max Kaiser
Austrian National Library
Josefsplatz 1, A-1015 Vienna, Austria
Tel.: +43-1-53410-370, Email: max.kaiser@onb.ac.at

Andrew Lindley
Austrian Research Centers GmbH –ARC
Donau-City-Strasse 1, 1220 Vienna, Austria
Tel.: +43-50550-4272, Email: andrew.lindley@arcs.ac.at

Abstract

This paper presents the Planets Testbed, a web-based application that provides its users with a controlled collaborative environment for scientific experimentation in digital preservation. The paper gives an overview about the core concepts of the Planets Testbed and describes how the application supports the user community in preserving the digital cultural heritage.

Keywords: Planets project, Testbed, digital preservation, long term preservation

Introduction

The Planets Testbed is one of the core results of the FP6 Planets Project (<http://www.planets-project.eu>) which aims to create a software suite capable of addressing the digital preservation challenges that libraries, archives and the digital preservation community are currently facing.

The Planets Testbed is more than a software package – it is a central environment (consisting of software, hardware and data) for testing the performance and capabilities of tools for digital preservation. The tools are offered as web services which can be combined in complex workflows. Measurement processes are highly automated, allowing large amounts of tool evaluation results to be collected via mass experimentation.

The Planets Testbed is essentially community software dedicated to people dealing with long term preservation issues on a day-to-day basis. In the following, we will provide an overview

of the Planets Testbed and discuss its role for the dedicated user community and for the preservation of the digital cultural heritage.

1. The Planets Testbed

1.1. The Planets Testbed Environment

The Planets Testbed provides a web-based software allowing to explore and test preservation services. This software relies on a Planets-wide, interoperable infrastructure, through which different tools can be invoked in a uniform way: the Planets Interoperability Framework. It defines the generic interfaces enabling the seamless integration of a large number of tools each of which provides a specific functionality required for performing long term preservation tasks.

1.2. The Experiment Process

Different kinds of experiments are divided into different 'Experiment Types' (see section 1.3). Each experiment type of is based on a workflow which itself consists of a sequence of preservation service operations.

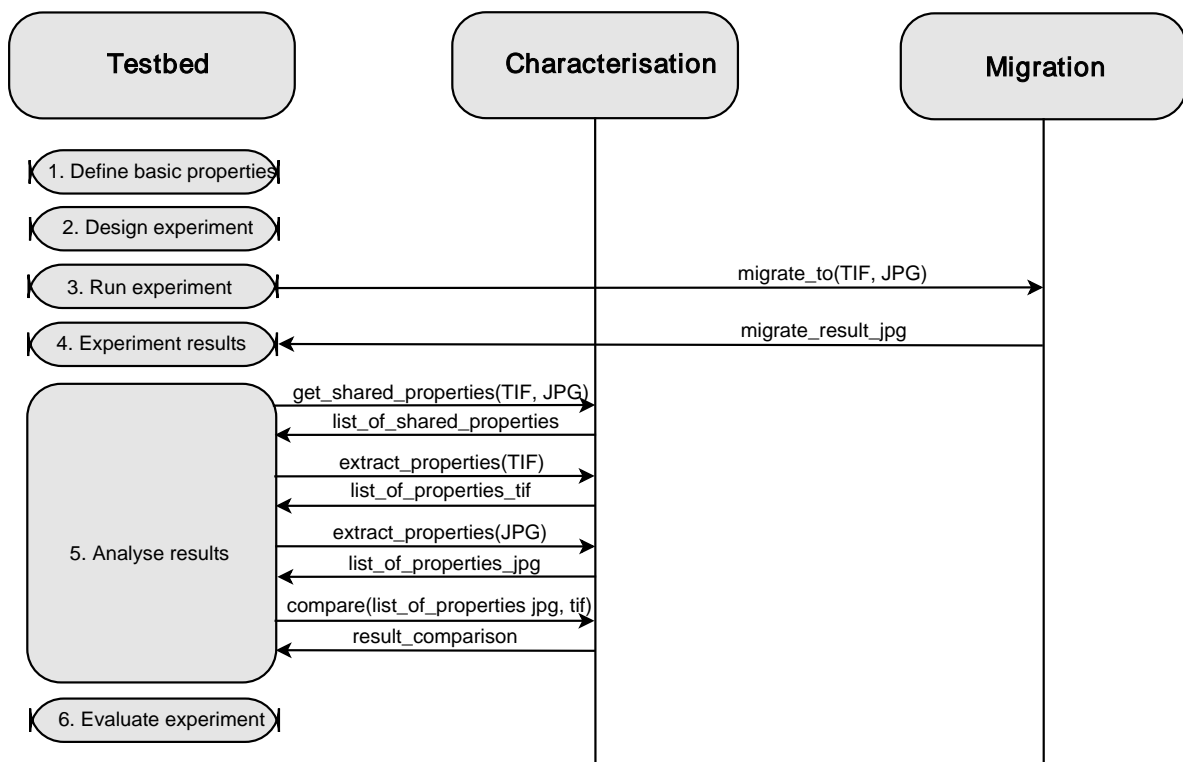


Figure 1: Example of a Planets Testbed experiment process

Using the Planets Testbed web application, the user is guided through six steps of an experiment process, as shown on the left-hand side of Figure 1. The following walk-through will use an example which might play a role in a real institutional process: The automated characterisation and migration of digital content. To be more concrete, this example refers to

the migration of a single TIF file to a single JPG file, and subsequently the comparison of the properties of the input and output files.

1.2.1. Define Basic Properties

In the first step of an experiment, basic experiment metadata is recorded. A user is required to enter a name for the experiment along with some basic information about the experimenter. The user can also supply information on the overall purpose and focus of the experiment, and references to relevant experiments, scientific publications or web resources.

1.2.2. Design Experiment

The experiment type can be selected here. A simple graphical representation of the experiment workflow is presented to the user. Configuration of this workflow depends on the experiment type, but in most cases, this involves browsing and selecting available services and selecting digital objects to experiment upon. The digital objects can be chosen from the data sets available in the Testbed or from content the user has uploaded.

Taking the example of the migration experiment, the workflow is configured by selecting a migration pathway, composed of the starting format TIF, the target format JPG, and a migration service (e.g. ImageMagik).

1.2.3. Run experiment

Once designed and configured, the experiment can be submitted for approval. At this point, the administrator in charge of the Planets Testbed is given an opportunity to prevent the experiment from being executed, for example if it is likely to put an unreasonable load on the server if executed at that time. Experiments that require only modest resources are automatically approved, and can be executed straight away.

Following approval, the user can initiate execution of the workflow. The Planets workflow execution engine then takes each digital object, and passes it through the specified chain of services.

1.2.4. Experiment results

In this step, the user can inspect the experiment result objects, overall success rates and basic performance statistics, e.g. whether all migration actions successfully created new digital objects. The user is also given the opportunity to re-run the experiment in order to collect additional data.

1.2.5. Analyse results

If characterisation tools are available for the digital objects which are part of the experiment, they can be used to analyse the properties of the digital objects. In our migration example, there are two digital objects, an input TIF file and the resulting JPG file which have different file format specific characteristics. Based on the common set of properties of these file

formats which are determined by a Planets characterisation service, the values can then be automatically compared using the metrics that apply to the different properties.

1.2.6. Evaluate Experiment

The final step of an experiment allows the user to judge the overall performance of the preservation workflow. The experimenter can also provide a brief written report about the experiment's outcome. The result can then be more widely shared between Planets Testbed users, so that others can learn from the results or even setup an equivalent experiment in order to reproduce and verify the outcomes of other experimenters.

1.3. Planets Testbed experiment types

An experiment type defines the generic structure and data flow of an experiment, and there are many kinds of experiments to be explored other than the migration experiment outlined as an example above. In the following, we shortly describe the experiment types that exist so far.

- *Characterisation Experiments*
A characterisation experiment allows for direct comparison of characterisation tools against each other or against a set of authoritative property values.
- *Validation Experiments*
A validation experiment is used to test whether a digital object is well-formed and valid with respect to a particular format.
- *Emulation Experiments*
Emulation generally refers to imitating a (usually obsolete) soft- and hardware environment within another (usually up to date) soft- and hardware environment. In the Testbed, an Emulation experiment creates an emulation session for a digital object which is then visualised the imitated soft- and hardware environment. By that way, the user can record how well the object is being rendered with respect to this specific environment.
- *Execute Plato preservation plan*
“Plato” (see [1]) is one of the outcomes of the Planets project, and is a web based software for creating a preservation plan for preserving a specific collection or a part of a collection of digital objects. The concrete recommendation of the preservation plan ends up in an “executable preservation plan” which can then be evaluated by a corresponding Planets Testbed experiment.

It is to be expected that the existing experiment types do not cover all the requirements for the different experiment scenarios the long term preservation community might require. If an experiment does not fit with one of the existing experiment types, a new experiment type must be set up by a Testbed administrator contacted through the Testbed helpdesk (see end of section 3).

2. Sharing knowledge with the Planets Testbed community

The Planets Testbed is community software in the sense that it allows reviewing and even reproducing existing experiments by all community members. New experiments can reference existing ones and refine or give a statement on existing experiment results. In that way the community members contribute to a continuously growing and reliable knowledge base on digital preservation.

The main goal of the Planets Testbed in this aspect is to enable community members to share their research results amongst cultural heritage institutions all over Europe. The Planets Testbed acts as the central experimentation platform gathering knowledge about long term preservation topics in various dimensions: In the first place, an experiment can focus on performance and reliability of long term preservation services and the underlying software components themselves. Then, the annotated experiment datasets contain information about special cases (an extreme value for a file format specific parameter, for example) and important properties of digital objects. And finally, the Planets Testbed establishes a procedure to share meaningfully aggregated results with other Planets software, like Plato (see [1]), for example.

2.1. Knowledge about long term preservation services

A wide range of preservation services have been developed by the Planets project, and the Planets Testbed aims to make them available for public use. Each service is supplied with metadata describing the supported formats, migration pathways, the identity of the service creator, the location of the endpoint which makes the Planets service available and so on. The Planets Testbed makes it easy to explore this information which is continuously managed and maintained.

2.2. Knowledge about experiment datasets

Some experiment types require information about the data an experiment is based upon. The Planets Testbed integrates annotated datasets (corpora) in order to be able to check the output of a service against recorded metadata. As a simple example, if an identification tool is tested against an object of a known format (e.g. PDF file), the Planets Testbed can compare the embedded properties against the results from the identification service. This allows the scope and accuracy of identification tools to be closely examined. Similarly, validation services can be exercised using carefully constructed valid and invalid documents, testing the edge-cases of format specifications. For example, the Isartor test suite (<http://www.pdfa.org/doku.php?id=pdfa:en:isartor>) can be used to detect whether validation tools can spot PDFs that are invalid with respect to the PDF/A-1 (ISO 19005-1:2005) specification.

2.3. Contributing to the Planets-wide knowledge base

By standardising and sharing results, the Planets Testbed acts as a central point for accumulation and aggregation of data from many experiments and across institutional boundaries. From this rich dataset it should be possible to determine the robustness and performance of particular preservation tools and techniques in an objective manner.

The results are stored centrally and can be used as a basis for future development of a knowledge base.

3. The Public Planets Testbed

The Planets Testbed software will be made publicly available by the Planets project. A full installation requires all of the different preservation services to be installed, each of which may have different software dependencies and operating system requirements. The publicly available central Planets Testbed addresses this problem by providing as many tools and services as possible – pre-installed, configured and ready for testing. The Planets Testbed can be accessed using a web browser and allows interested parties to evaluate all the preservation services and strategies supported by Planets using their own data or benchmark content.

Additionally, it is possible to download and install individual Planets Testbed instances. The software installer makes it easy to deploy the Planets Testbed locally, but can only provide limited functionality out of the box.

The public Planets Testbed is available at <http://testbed.planets-project.eu/testbed>, hosted by HATII at the University of Glasgow. It is currently in beta release phase and selected external parties have accounts granted. The service will go completely public in beginning of 2010, but it is already possible to ask for an account at helpdesktb@planets-project.eu. Further information about the Planets Testbed, also about upcoming training workshops can be found on the Planets website.

Conclusions

The innovative aspects of the Planets Testbed are the ways in which experimental data is collected, analysed and shared. The Planets Testbed provides a single interface to a wide range of hardware and software benchmarking environments, so that data can be collected reliably and reproducibly.

The Planets Testbed is also building corpora of digital objects with well-known properties. These properties, in combination with a number of innovative Planets software technologies, allow for the outputs of preservation services to be analysed rigorously and automatically.

Finally, the Planets Testbed defines standard semantic structures to contain these results, permitting community-wide aggregation of experimental results and experiences using the tools and services needed for long-term preservation of the digital cultural heritage.

The Planets Testbed will be made available to the digital preservation community as a free service by beginning of 2010.

References

1. Christoph Becker, Hannes Kulovits, Andreas Rauber, and Hans Hofman, Plato: a service oriented decision support system for preservation planning, JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries (New York, NY, USA), ACM, 2008, See <http://doi.acm.org/10.1145/1378889>.



1378954, pp. 367-370.

2. Petra Helwig, Judith Rog, Caroline van Wijk, Eleonora Nicchiarelli, and Manfred Thaller, Test methods for testbed, Tech. report, 2007, See http://www.planets-project.eu/docs/reports/Planets_TB3-D2_MethodsForTesting.pdf.